INTRODUCTION OF RNA FOLDING AND HIGHER-ORDERED STRUCTURE

Shu-Yun Le
Lab. of Experimental and Computational Biology
NCI Center for Cancer Research, NCI, NIH

We know a number of important roles for RNA structure

Catalytic(e.g. ribosomal RNA is added to the list of ribozymes.)

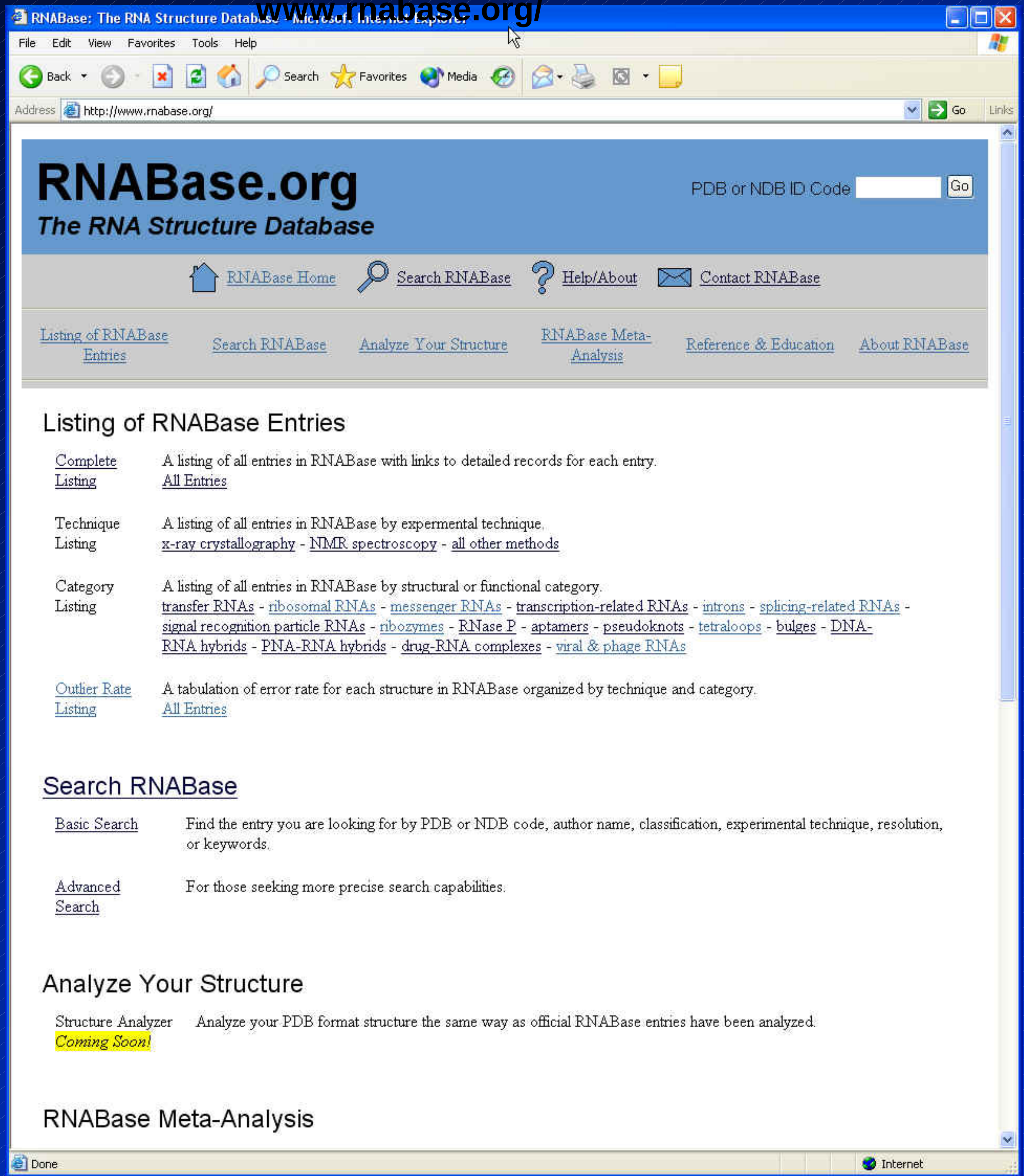
Binding(RNA-protein, RNA-RNA, RNA-DNA and RNA-small molecule, other?) involved in regulation and localization.

Purely structural?

RNA Informatic issues

Some databases

Analysis program repositories and servers www.rnabase.org/



Address http://www.imb-jena.de/RNA.html

www.imb-jena.de/RNA.html

imb Institut für Molekulare Biotechnologie JENA

The RNA World Website

Databases, Web Tools | Software | Online Books and Tutorials | Meetings | Miscellaneous | Search

Welcome to ***The RNA World Website*** at IMB Jena. This web resource lists Internet links on RNA related topics.

- Have a look at a short article describing this site: J. Sühnel, *Trends in Genetics* *1997, 13, 206-207*, Views of RNA on the World Wide Web (reprint version in PDF format, PubMed link).
- Read a WebWatch description of this website in *Nature Reviews: Molecular Cell Biology* *2002, 3, 3-9.* [WebWatch is on p. 4; PDF].
- The RNA World Website has been included in the Web 's Best Sites collection of the *Encyclopedia Britannica*.

- Breakthrough of the Year: 2002 (20 December 2002 issue of Science)

- D. Kennedy, Editorial, Science **298**, 2283 (2002)
- J. Couzin, Breakthrough of the Year: Small RNAs Make Big Slash, Science **298**, 2296 (2002)

Science is making the full text of the online edition of the 20 December 2002 Science, including the Breakthrough of the Year section, available free of charge to all registered users of Science Online.

Databases, Web Tools

Three-dimensional structures (coordinates and images)

- *The Nucleic Acid Database (NDB)*
- *The Protein Data Bank (PDB)*
- *The RiboWeb Project (three-dimensional models of the E. coli 30 S ribosomal subunit and 16 S rRNA)*
- *RNase P 3D models*
- *IMB Jena Image Library of Biological Macromolecules*
  - (with a compilation of all RNA structures from the Protein Data Bank)
- *The RNA Structure Database*
- *SCOR: Structural Classification of RNA*
- *Visualization of Viruses (DNA and RNA)- University of Wisconsin, Madison*
- *Ribosome Images (Wadsworth Center Microscope 3D Database)*
- *Base pairs*
  - *Compilation by Tinoco*
  - *Compilation by Dirheimer et al.*
  - *Database of non-canonical base pairs found in known RNA structures (Fox Lab)*
  - *RNA base pair isostericity (Leontis, Westhof)*
  - *The Base Pair Directory of the IMB Jena Image Library of Biological Macromolecules*

Sequences, Secondary structures, Other

- *5S Ribosomal RNA Database*
- *Database of Ribosomal Crosslinks (DRC)*
- *Ribosomal Database Project II*
- *Ribosomal RNA Mutational Database*
- *European Large Subunit Ribosomal RNA Database*
- *European Small Subunit Ribosomal RNA Database*
- *Ribosomal Internal Spacer Sequence Collection (RISSC)*
- *Comparative RNA Web Site*
  - *Old RNA Secondary Structures Site*

- *tRNA and tRNA Gene Sequences*
- *GtRDB: The Genomic tRNA Database*
- *PLMItRNA: A Database for Plant Mitochondrial tRNA Genes and Molecules*
- *Aminoacyl-tRNA Synthetases Database (AARS)*

- *tmRNA Database*
- *tmRNA Website*

RNA Structures

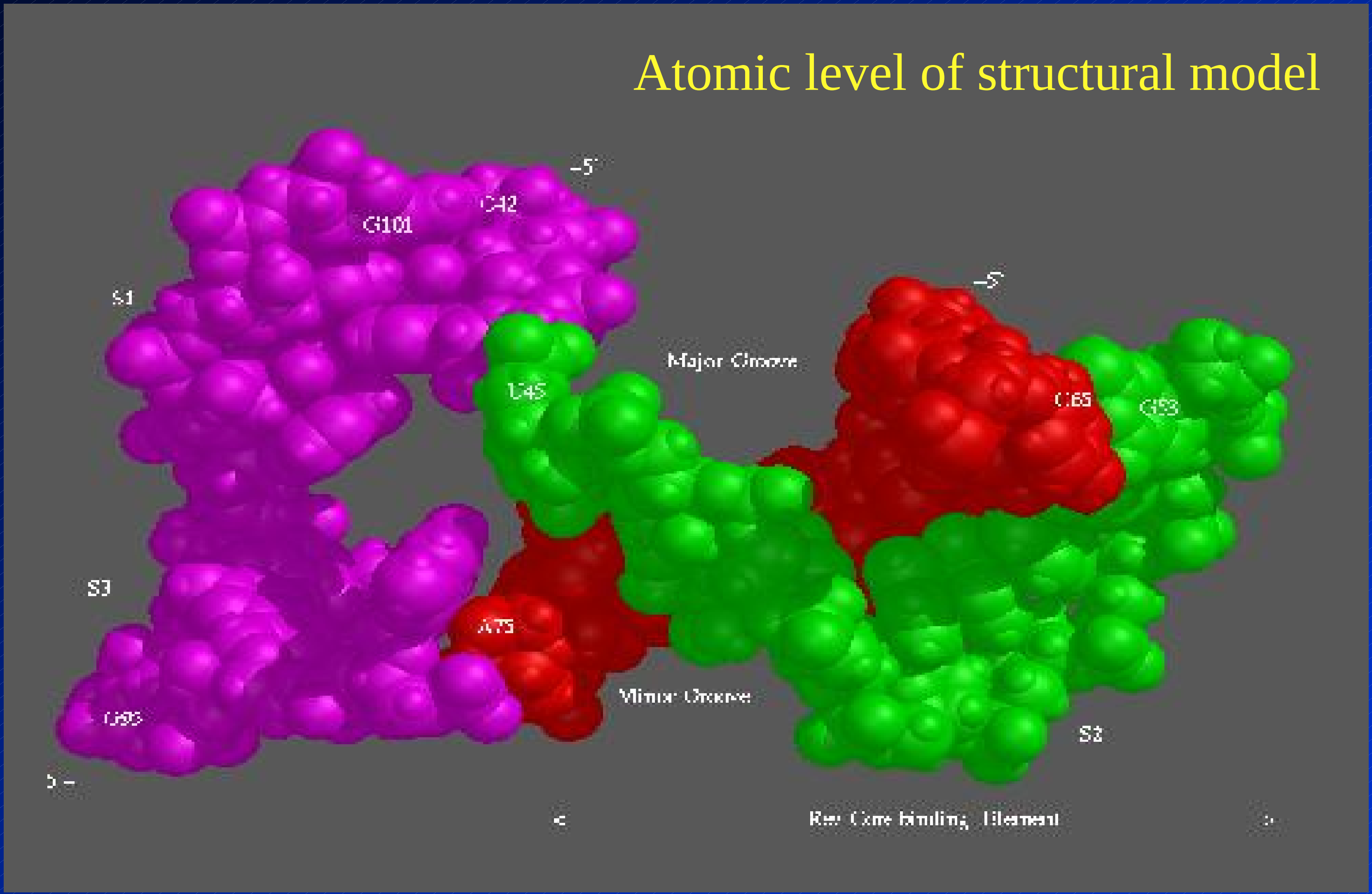
Primary sequence

Secondary structure

Tertiary structure(e.g. pseudoknots, etc.)

Ultimately atomic scale models

Three dimensional model of the core structure of HIV-1 RRE



RNA SECONDARY STRUCTURE

Well-ordered secondary structure required for RNA function

- Ribozymes
- Ribosomes
- Signal recognition particle (srp) RNAs
- transfer RNAs (tRNAs)
- Functional RNA elements
  - RRE, TAR, IRES, IRE

Some Rules for RNA Folding (severely simplified)

Single stranded nucleic acids can fold back on themselves to form regions of typical duplex structure(called “stems”)

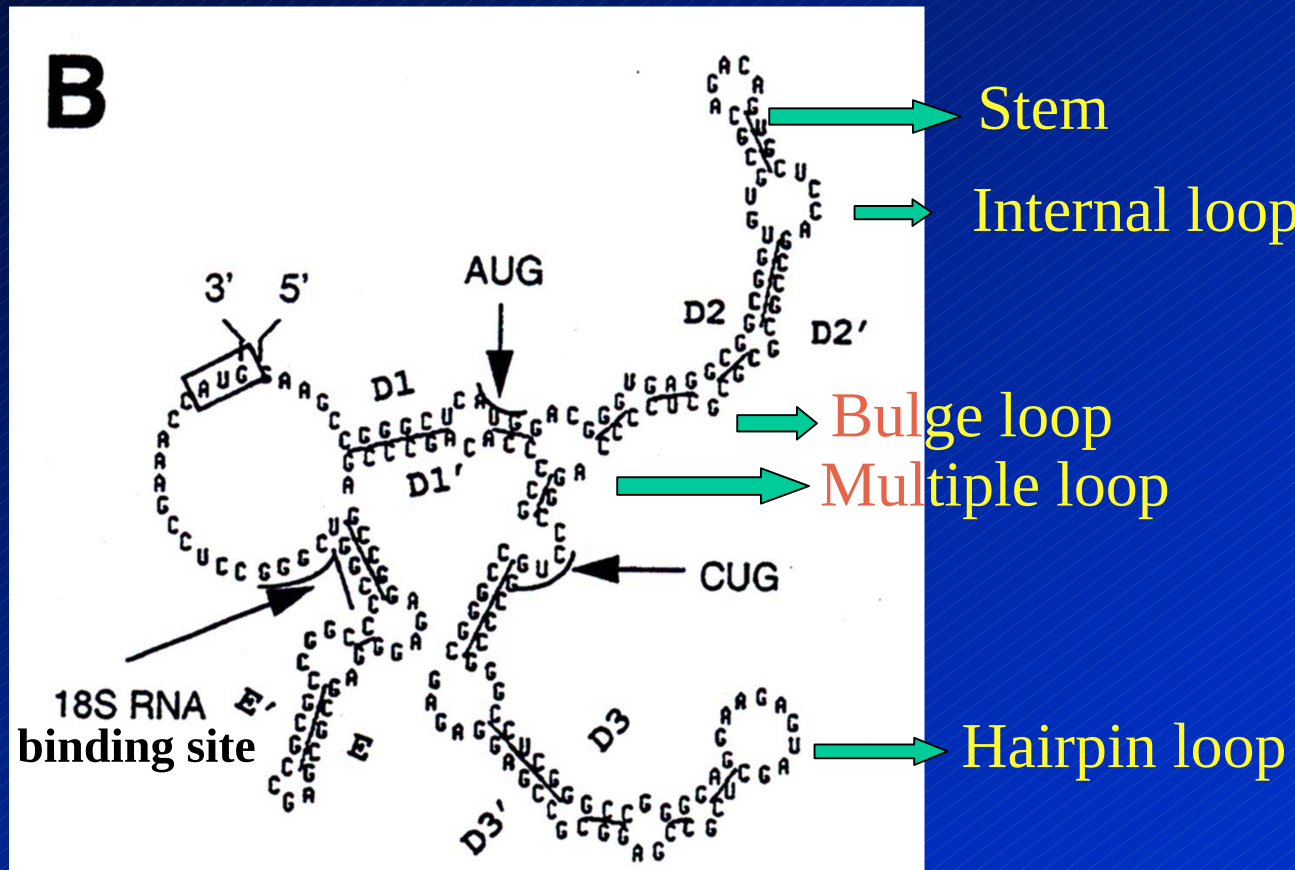
Watson-Crick rules: **A:U, G:C, (G:U -”wobble”)are favorable**

Thermodynamic refinements:

- **benefits for helical stacking**
- **penalties for loops (hairpin, internal, and multi-branched).**

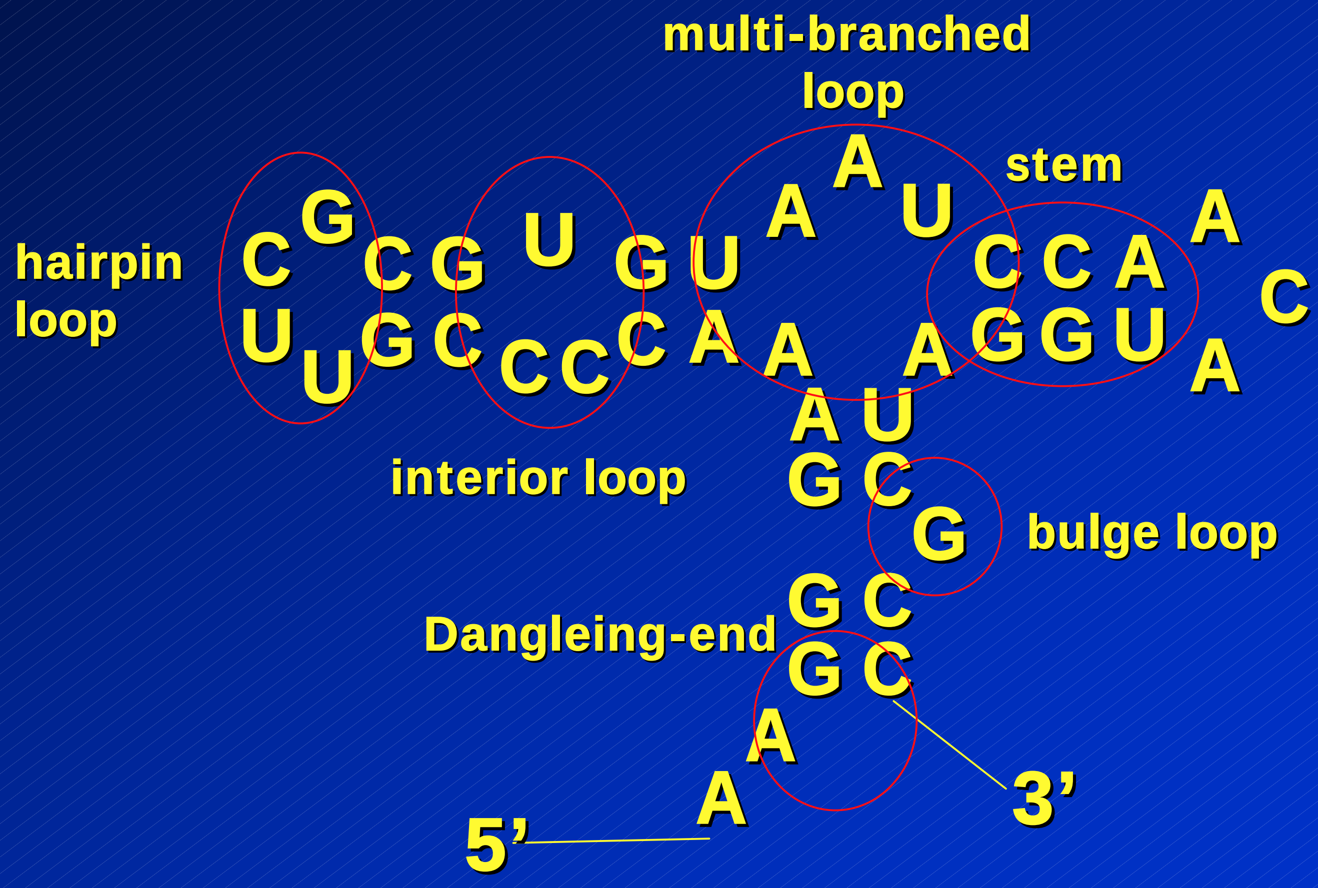
Rules are undergoing long term, gradual refinement so that they can now correctly predict “most” of the base pairs observed in known structures.

An example of RNA secondary structure

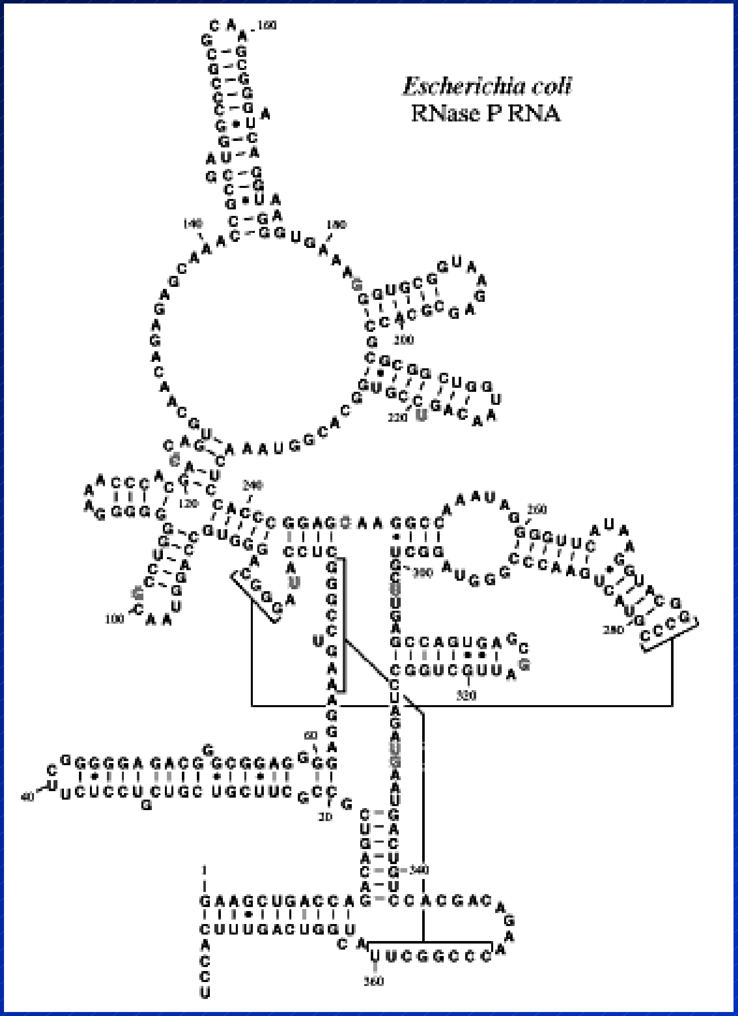


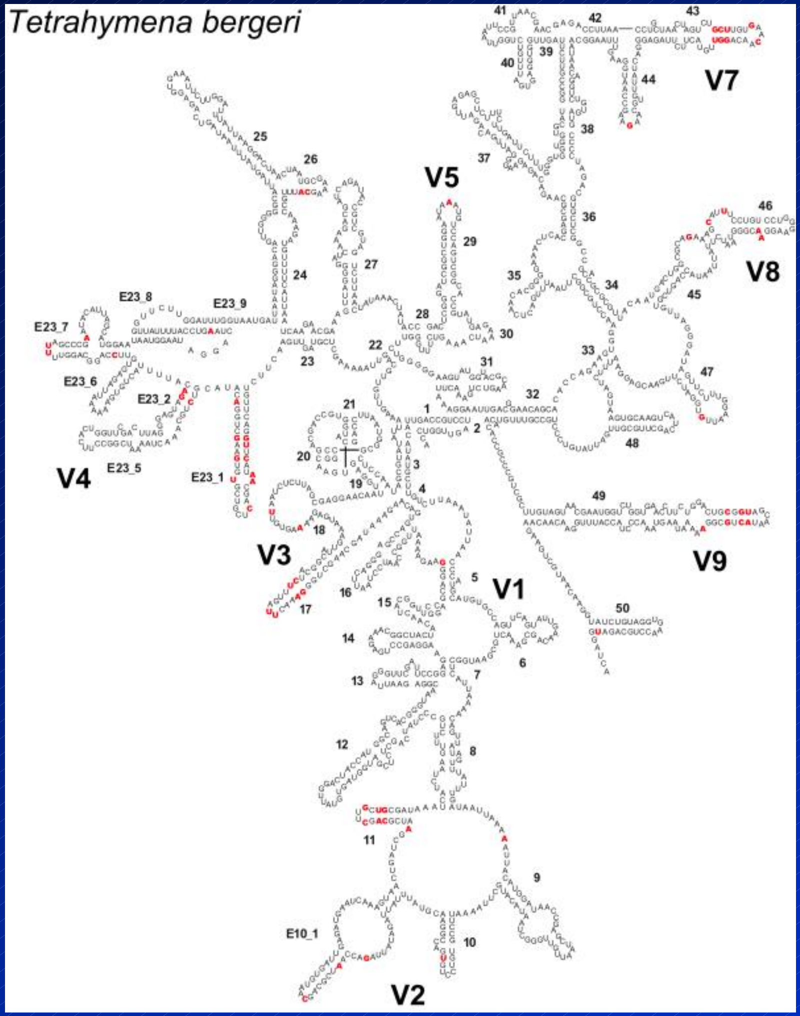
RNA SECONDARY STRUCTURE

Nomenclature



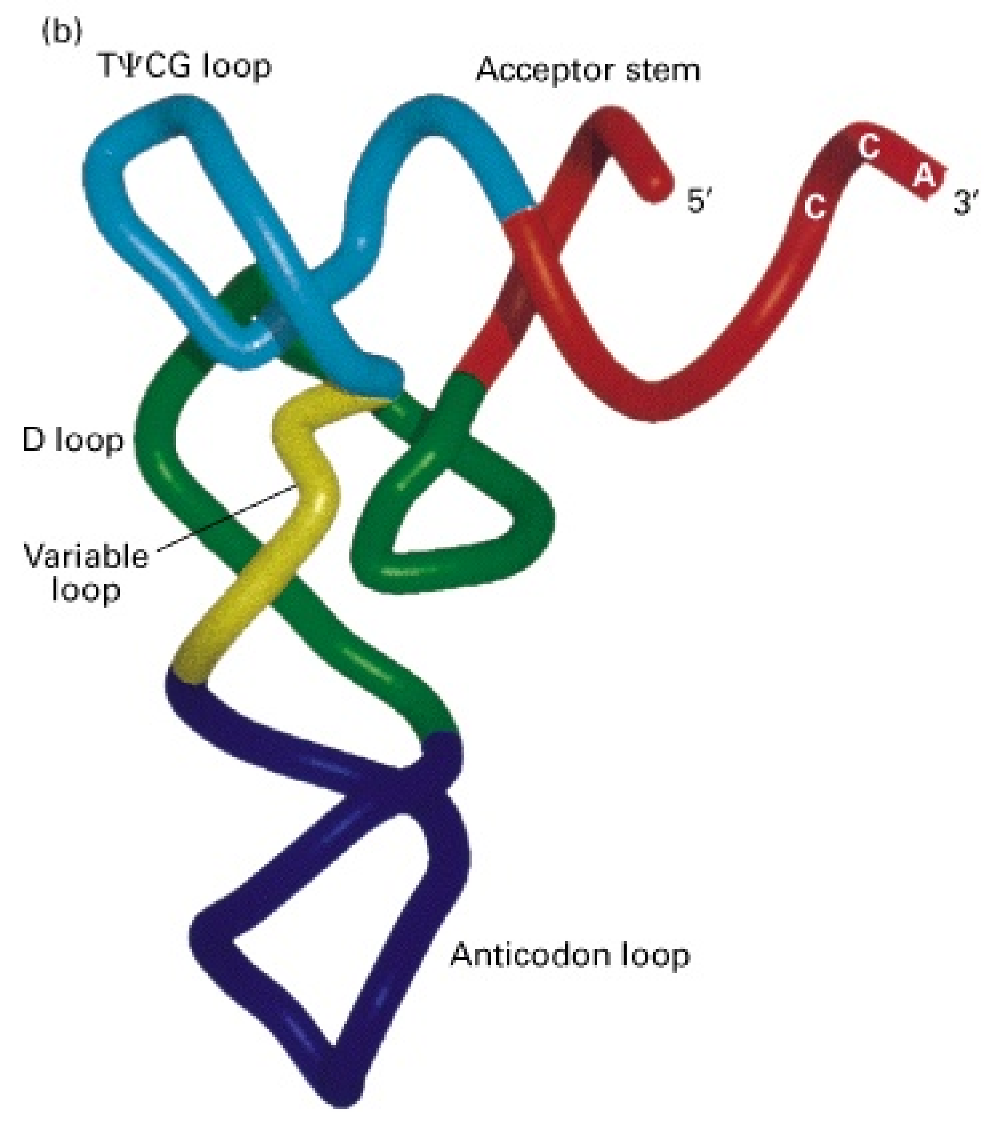
RNA SECONDARY STRUCTURE
Of RNase P RNA





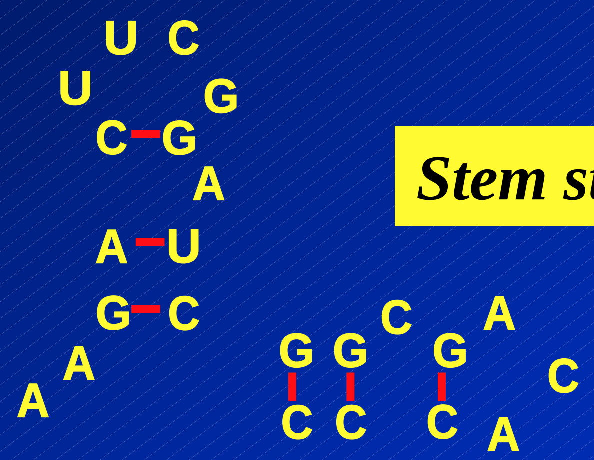
Small subunit ribosomal RNA

transfer RNA tertiary structure



RNA SECONDARY STRUCTURE

Stems are nested relationships



Stem stacking

AAGACUUCGGAUCUGGCGACACCC

RNA SECONDARY STRUCTURE

Stems are nested relationships

If positions i and j pair and i’ and j’ pair, these pairs are nested if:

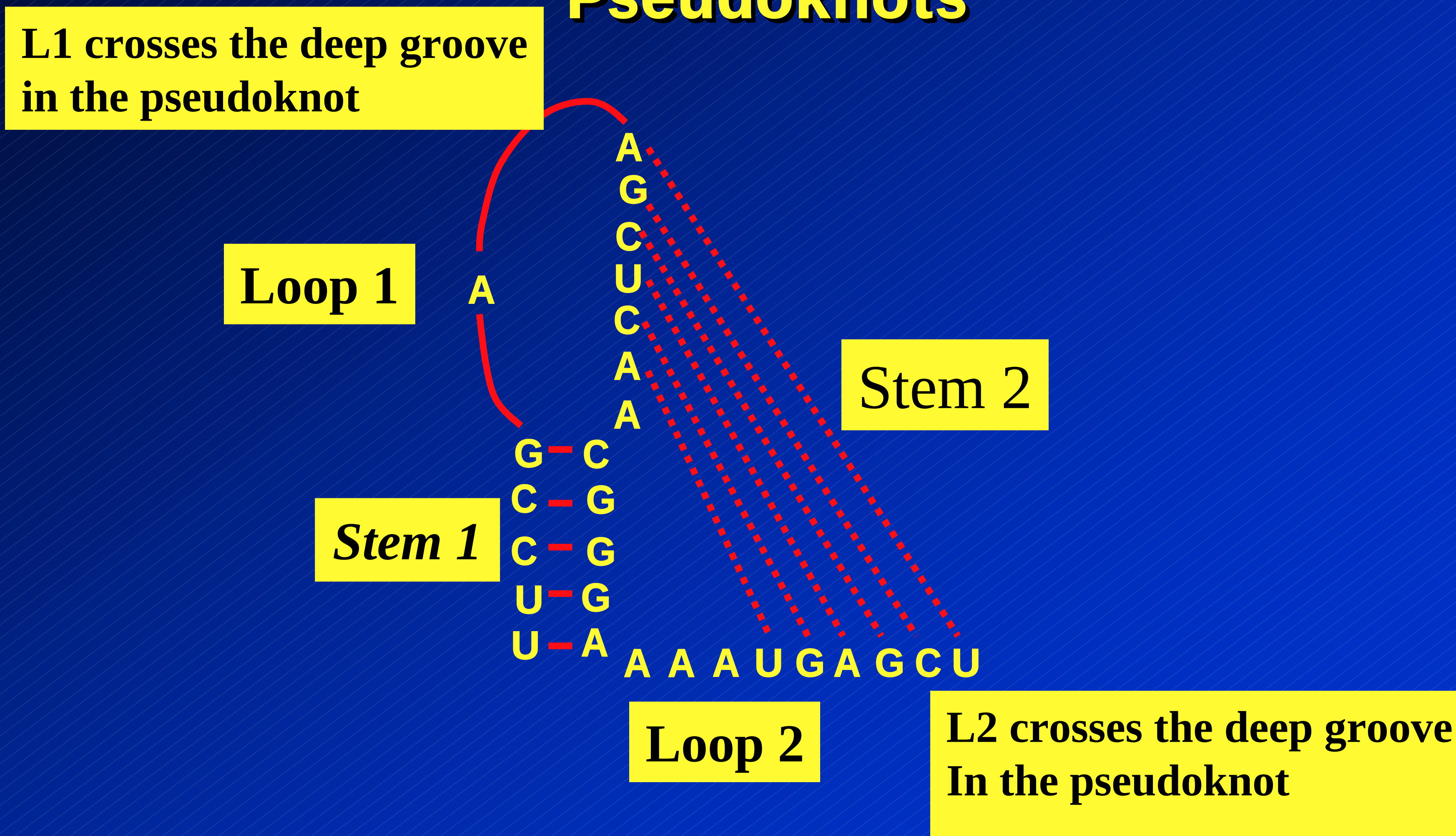
$$i < i' < j' < j \quad \text{or} \quad i' < i < j < j'$$

AAGACUUCGGAUCUGGCGACACCC

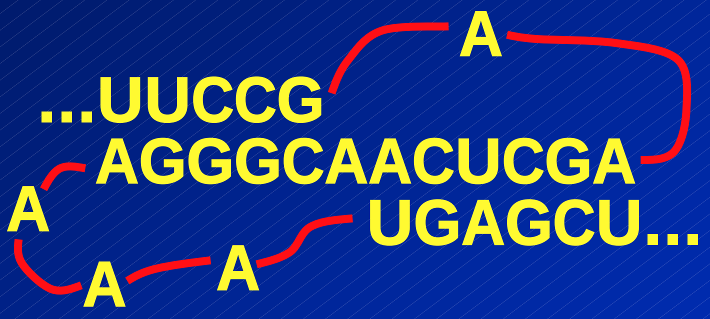
3 5 10 13 15 18 22 24

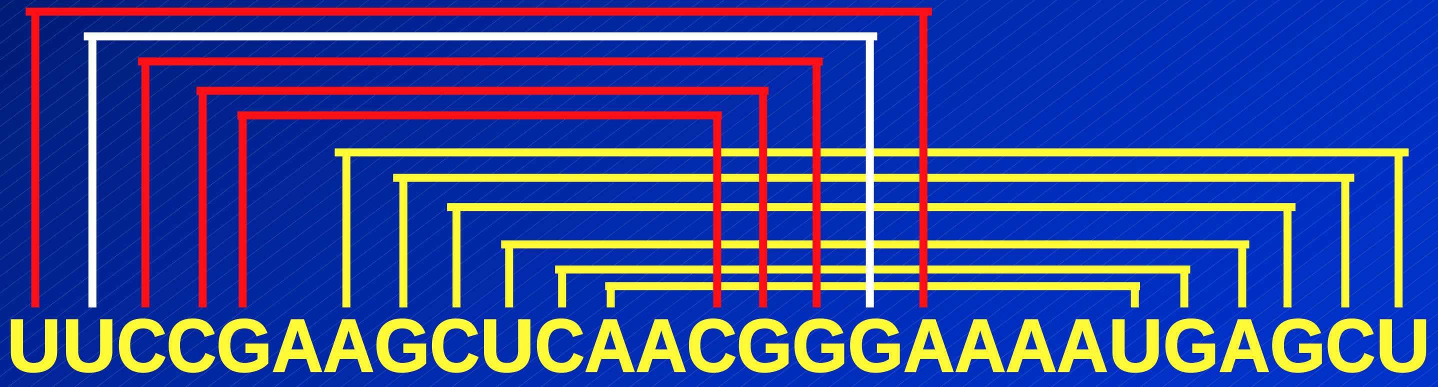
RNA TERTIARY STRUCTURE

Pseudoknots



Pseudoknots are not nested





RNA SECONDARY STRUCTURE

Predicting for secondary structures

Comparative sequence analysis

Nussinov folding algorithm

Zuker folding alogrithm

Genetic Algorithm (RNAGA, Chen, Le and Maizel)

PREDICTION of RNA SECONDARY STRUCTURE

Comparative sequence analysis

Problem: multiple solutions, very tedious manual works are often involved.

Computation: Search for conserved complementary base-pairings in the folded stems.

Requirement: A set of conserved phylogenetic sequences and a reasonable multiple sequence alignment.

PREDICTION of RNA SECONDARY STRUCTURE

Comparative sequence analysis:

Step 1: Multiple sequence alignment

Step 2: Search for covarying nucleotides

```
Seq1  GCCUUCGGGC
Seq2  GACUUCGGUC
Seq3  GCCUUCGGGC
```

```
 U C
U   G
 C-G
 N-N'
 G-C
```

PREDICTION of RNA SECONDARY STRUCTURE

Comparative sequence analysis:

Measuring pairwise sequence covariation

Mutual Information: A measure of how much uncertainty about the nucleotide at one site is reduced by knowing the nucleotide at another site.

PREDICTION of RNA SECONDARY STRUCTURE

Comparative sequence analysis:

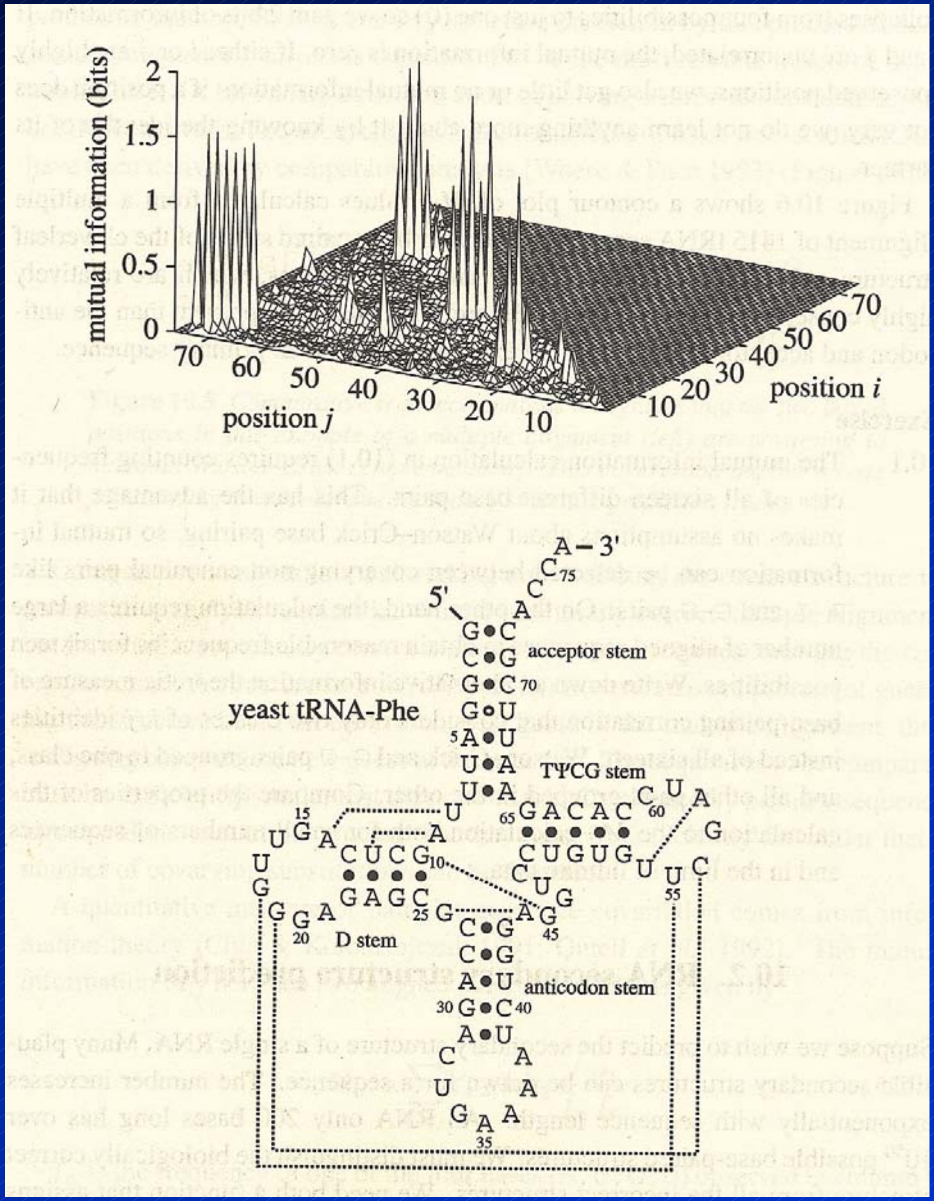
Mutual Information:

$$\mathrm{MI}(X,Y) = \sum_i \sum_j \mathrm{P}(X_i, Y_j) \log_n \frac{\mathrm{P}(X_i, Y_j)}{\mathrm{P}(X_i)\mathrm{P}(Y_j)}$$

$P(X_i)$ is the probability (frequency) of nucleotide i at site X

$P(X_i,Y_j)$ is the joint probability of nucleotide i at site X and nucleotide j at site Y

RNA SECONDARY STRUCTURE



From: Durbin, et al. 1998. Biological Sequence Analysis.

PREDICTION of RNA SECONDARY STRUCTURE

Nussinov RNA folding algorithm:

Dynamic programming algorithm for finding the RNA secondary structure with the maximum base-pairs.

Recursive algorithm, building larger subsequences onto smaller subsequences.

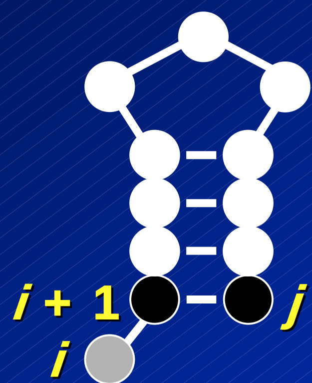
For nucleotides *i* and *j* there are 4 ways to add them to an existing subsequence structure:

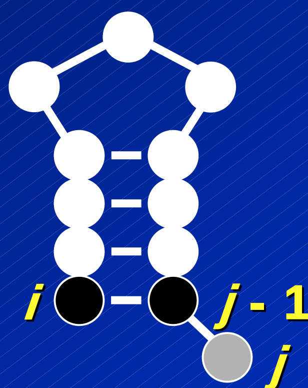
- Add unpaired *i* to structure for i+1, *j*
- Add unpaired *j* to structure for *i*, *j*-1
- Add *i*,*j* pair to structure for *i*+1,*j*-1
- Combine 2 optimal substructures *i*,*k* and *k*+1, *j*

PREDICTION of RNA SECONDARY STRUCTURE

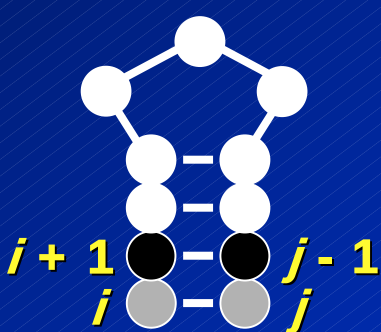
Nussinov RNA folding algorithm

i unpaired

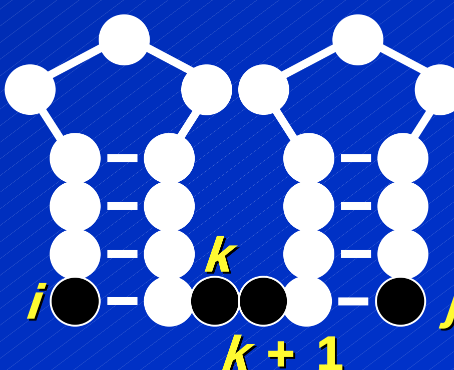




j unpaired



ij pair



bifurcation

PREDICTION of RNA SECONDARY STRUCTURE

Nussinov RNA folding algorithm

The algorithm:

Given a sequence x of L symbols $x_1, \ldots, x_L$

If x_i and x_j are complementary base pairs then $\delta(i,j) = 1$

If not, then $\delta(i,j) = 0$

Calculate scores $\gamma(i,j)$ that are the maximal number of base pairings that can be formed for subsequence $x_i \ldots x_j$

As with other dynamic programming alogrithms we've seen there are two steps: filling the matrix followed by traceback.

PREDICTION of RNA SECONDARY STRUCTURE

Nussinov RNA folding algorithm

Matrix filling

Initialize the matrix $\gamma(i, i-1) = 0$ for i = 2 to L

$\gamma(i, i) = 0$ for i = 1 to L

Recursive filling for subsequences of length 2 to L

$$\gamma(i,j) = \max : \begin{cases} \gamma(i+1, j) \\ \gamma(i, j-1) \\ \gamma(i+1, j-1) + \delta(i,j) \\ \max [\gamma(i, k) + \gamma(k+1, j)] \\ i < k < j \end{cases}$$

PREDICTION of RNA SECONDARY STRUCTURE

Nussinov RNA folding algorithm

Matrix filling

	A	G	G	G	A	A	A	A	U	C	C	C
A	0	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**2**	**3**	**4**
G	0	0	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**2**	**3**	**4**
G		0	0	**0**	**0**	**0**	**0**	**0**	**1**	**2**	**3**	**3**
G			0	0	**0**	**0**	**0**	**0**	**1**	**2**	**2**	**2**
A				0	0	**0**	**0**	**0**	**1**	**1**	**1**	**1**
A					0	0	**0**	**0**	**1**	**1**	**1**	**1**
A						0	0	**0**	**1**	**1**	**1**	**1**
A							0	0	**1**	**1**	**1**	**1**
U								0	0	**0**	**0**	**0**
C									0	0	**0**	**0**
C										0	0	**0**
C											0	0

Delta matrix

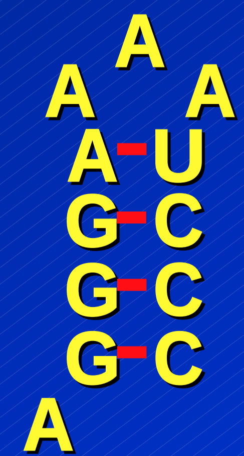
	A	G	G	G	A	A	A	A	U	C	C	C
A	0	0	0	0	0	0	0	0	1	0	0	0
G	0	0	0	0	0	0	0	0	0	1	1	1
G	0	0	0	0	0	0	0	0	0	1	1	1
G	0	0	0	0	0	0	0	0	0	1	1	1
A	0	0	0	0	0	0	0	0	1	0	0	0
A	0	0	0	0	0	0	0	0	1	0	0	0
A	0	0	0	0	0	0	0	0	1	0	0	0
A	0	0	0	0	0	0	0	0	1	0	0	0
U	0	0	0	0	1	1	1	1	0	0	0	0
C	0	1	1	1	0	0	0	0	0	0	0	0
C	0	1	1	1	0	0	0	0	0	0	0	0
C	0	1	1	1	0	0	0	0	0	0	0	0

PREDICTION of RNA SECONDARY STRUCTURE

Nussinov RNA folding algorithm

Traceback

	A	G	G	G	A	A	A	A	U	C	C	C
A	0	0	0	0	0	0	0	0	1	2	3	4
G	0	0	0	0	0	0	0	0	1	2	3	4
G		0	0	0	0	0	0	0	1	2	3	3
G			0	0	0	0	0	0	1	2	2	2
A				0	0	0	0	0	1	1	1	1
A					0	0	0	0	1	1	1	1
A						0	0	0	1	1	1	1
A							0	0	1	1	1	1
U								0	0	0	0	0
C									0	0	0	0
C										0	0	0
C											0	0



PREDICTION of RNA SECONDARY STRUCTURE

Zuker Thermodynamic Programming algorithm

Energy minimization

Sum of contributions from: loops
base pairings
bulges
other sequence elements

And stacking interactions

For example: GC different from AU
GC GC

PREDICTION of RNA SECONDARY STRUCTURE

Zuker Thermodynamic Energy minimization:

```
4 nt loop +5.9          AA
                       A  A      -1.1 terminal mismatch
-1.8 stack              G-C
-0.9 stack              A-U
                        U-A
1 nt bulge  +3.3      A          -1.8 stack (bulge)
-2.9 stack              G-C
-2.9 stack              G-C
-1.8 stack              C-G
                        U-A
5' dangle -0.3         A
unstructured single   A
strand           0.0
```

Overall ΔG = -4.3 kcal/mol

PREDICTION of RNA SECONDARY STRUCTURE

Zuker Energy minimization

- Program: Mfold
- Uses dynamic programming to predict secondary structures in RNA sequences
- Returns optimal and suboptimal predicted structures
- Dynamic programming requires 2 matrices, V and W
- W(i,j) := energy of best structure on i,j
- V(i,j) := energy of best structure on i,j given that i,j are paired

PREDICTION of RNA SECONDARY STRUCTURE

RNAGA: A genetic algorithm for predicting a secondary structure common to a set of phylogenetically related sequences.

Algorithm: RNA structure is optimized by not only the free energy of the formation of the structure but also the structural similarity among the homologous sequences by a genetic algorithm.

Genetic Algorithm of RNAGA:

1. **A population of individuals (a set of RNA secondary structures)**
2. **A measure that provides the fitness for the individuals**
3. **Operations that are intended to model crossover, mutation and selection.**

PREDICTION of RNA SECONDARY STRUCTURE

RNAGA Algorithm:

Individual Representation:

A secondary structure is an individual in the population. Each structure is encoded by *T* as a set of Stems,

$$T = \{S_1, S_2, \ldots, S_n\}$$

$S_i = (a_i, b_i)$, (a_i, b_i) is the closing base pair of a stem S_i

Each of individuals represents a search point in the space of potential solutions to a given optimization based on the selected *Fitness Function*

PREDICTION of RNA SECONDARY STRUCTURE

RNAGA Algorithm:

Fitness Function

Predefined fitness (object) functions:

Thermodynamic Stability: Folded Free Energy

It is used in the initial stage of the optimization procedure

Structural Similarity among the structures:

It is used in the second stage of the optimization.

PREDICTION of RNA SECONDARY STRUCTURE

RNAGA Algorithm:
Initial Generation of a *Population of Individuals*

For each sequence

a. A stem S_i is randomly chosen from the *master list* of all possible stems that can be formed by base-pairing rule.
b. For the stem S_i we consider a list of stems that are interior (nested) to the stem. From this list we select those stems that are compatible with those already incorporated into the structure until no stem can be added.
c. In the procedure, a stem is added to the structure if the addition of a stem increases the stability of the structure, otherwise, the addition is determined by the Boltzmann rule.
d. The steps a-c are repeated until no more stem S_i can be chosen from the *master list*

PREDICTION of RNA SECONDARY STRUCTURE

RNAGA Algorithm: *Operation*

Crossover: Genetic crossover exchanges information among solutions creating the possibility of the right combination of motifs for better solutions.

In *RNAGA,* a pair of structures is selected as two parental structures from the population and the selection is based on the fitness score. A stem pool is formed from the pair of structures. An offspring of the two parental structures is constructed by stepwise selection of one stem after another from the pool. Only stems compatible with the previously selected ones are added. If two stems overlap and one is selected then the selected one is taken wholly and the other is shortened.

PREDICTION of RNA SECONDARY STRUCTURE

RNAGA Algorithm: Crossover Operation:

The offspring is required to be different from the two parental structures.

For a population of *n* structures *n* pairs of structures are selected to be subjected to crossover

PREDICTION of RNA SECONDARY STRUCTURE

RNAGA Algorithm: Mutation Operation

Every structure in the population is subjected to be mutated.

Mutation is performed by the removal of some stems from the individual and the subsequent addition of new stems.

a. **In the initial stage, the stem which closes a region with positive *E* will be removed from the structure. If no such stem exists, the removal stem is randomly selected.**
b. **In the second stage, the removal of stems is based on a roulette wheel spin method with slots weighted in inverse proportion to the stem conservation scores. The addition of new stem is done in a randomized manner.**
c. **Resulted mutated structure is required to possess a certain thermodynamic stability (predetermined).**

PREDICTION of RNA SECONDARY STRUCTURE

RNAGA Algorithm: Selection Operation

a. For a population of *n* structures, *3n* structures are produced in each GA iteration.

b. Size of population is kept constant in the algorithm.

c. Fitness Scores:

Free Energy *E* , Structural Conservation (Similarity Score)

Stem Conservation, Structural Distance Function, d_{ij}

$d_{ij} = 1 - n_{ij} / m_{ij}$, **$n_{ij}$ is the number of base pairs in common between S_i and S_j; m_{ij} is the maximal of base pairs between two any structures of the population.**

d. Structure difference between solutions

For each structure we also define a score as the difference between its fitness and the best fitness value in the set divided by its distance function,

PREDICTION of RNA SECONDARY STRUCTURE

RNAGA Algorithm: Selection Operation

e. The structures in the population are sorted in increasing order of this score and the new population is selected from the top of the list.

Implementation of RNAGA

1. Generate an initial population of *n* structures for each Sequence.

2. Iterate crossover, mutation and selection with *E* as the fitness function until the stability criteria of the structures are reached. In the second stage, those operations are optimized based on the fitness functions of both E and structural similarity.

3. Evaluate the conservation score for each structure in the current generation and compute the stem conservation score for each stem in the structure for each sequence.

4. Perform genetic operations on the current generation for each sequence.

5. Collect potential common structures for each sequence.

Implementation of RNAGA

6. Select the next generation for each sequence.

7. Repeat steps 3-6 until the maximal number of generations has been reached and converge is reached.

8. Rank those structures based on the computed conservation scores.

Implementation of RNAGA

RNAGA Web Interface:

http://protein3d.ncifcrf.gov/shuyun/rnaga.html

RNAGA server for online users:

http://protein3d.ncifcrf.gov/shuyun/dorna2d.html

Accuracy of RNAGA predictions

Table 1. Accuracy of a genetic algorithm for RNA common secondary structure prediction

RNA	Nucleotides	Base pair	Correctly predicted base pair (%)			
			Rank 1	Rank 10	Best structure	Any structure
tRNA	1556	432	87.7 ± 12.4	81.2 ± 12.5	98.8 ± 2.7	99.8
5S rRNA	3004	910	95.3 ± 7.0	87.9 ± 7.3	98.6 ± 4.3	98.7

Only the first 10 ranked ordered structures were considered in assessing the accuracy. The accuracy was determined for: the structure ranked first (i.e. with highest adjusted conservation score); the structure ranked tenth; the single best structure of the first 10 ranked ordered structures (column 6); the base pairs correctly predicted in at least one structure (column 7). The accuracy was determined by counting correctly predicted base pairs. Standard deviations are given with the percentages to demonstrate the range of accuracy. Only tRNA and 5S rRNA are listed since there are no known standard structures in the RREs of HIV-1 and HIV-2.

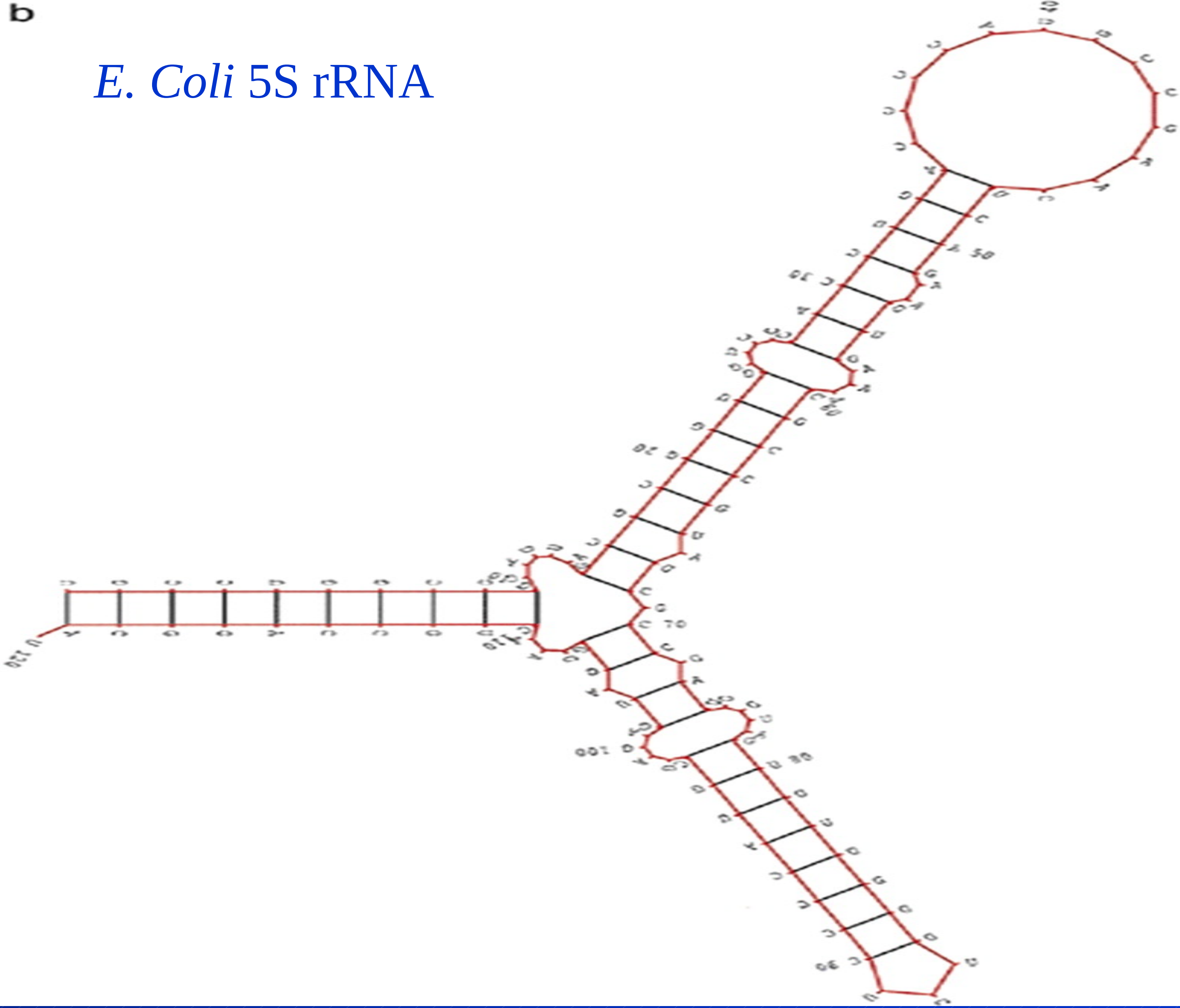
Alignment of 25 5S rRNAs

```
                      A                        B                                        B′
E.coli       UGCC.UGGCGGcaGUAGCGCGGUGguccCACCUGA.cccCAUGCcgaacUCAGaaGUGaaaCGC
Photobac     UGCU.UGGCGAccAUAGCGUUAUGgacccACCUGA.UCcCUUGCcGAacUCAGuaGUGaaaCGU
Beneckia     uGCU.UGGCGAccAUAGCGAUUUGgacccACCUGAcUUcCAUUCcGAacUCAGaaGUGaaaCGA
P. fluor     uUCU.UGACGAccAUAGAGCAUUGgaacCACCUGA.UCcCAUCCcGAacUCAGuaGUGaaaCGA
Azo.vine     UGCU.UGACGAUcAUAGAGCGUUGgaacCACCUGA.UCcCAUCCcGAacUCAGcaGUGaaaCGA
P.aerugi     uGCU.UGACGAUcAUAGAGCGUUGgaacCACCUGA.UCcCUUCCcGAacUCAGaaGUGaaaCGA
R.rubrum    UGGCC.UGGUGGUcAUUGCGGGCUCgaaaCACCCGA.UCcCAUCCcGAacUCGGccGUGaaaGAG
Wicrococ    G.UUA.CGGCGGCuAUAGCGUGGGGgaaacGCCCGG.cCGUAUAUCGaacCCGGaaGCuaagCCC
Streptom    G.uUU.CGGUGGUcAUAGCGUGAGGgaaacGCCCGG.UUaCAUUCcGAacCCGGaaGCuaagCCU
Th.aquat    AAUCCCCCGUGCCcUUAGCGGCGUGgaacCAC.CCGUUCcCAUUCcGAacACGGaaGUGaaaCGC
Th.therm    AAUCCCCCGUGCCcAUAGCGGCGUGgaacCAC.CCGUUCcCAUUCcGAacACGGaaGUGaaaCGC
Prochlor     uUCC.UGGUGUCuCUAGCGCuuuggaacCACuUCGAUUcCAUCCcGAacUCGAuuGUGaaacuu
Hb.salin     uu..aAGGCGGCcAUAGCGGUGGGguuacUC.CCGUaccCAUCCcgaacACGGaaGAuaagCCC
Hc.morrh     uu..aAGGCGGCcACAGCGGCGGGgcgacUC.CCGUaccCAUCCcgaacACGGcaGAuaagCCC
T.acidop     ...GGcAACGGUcAUAGCAGCAGGgaaacAC.CaGAUCcCAUUCcGAacUCgacGGUuaagCCU
S.acidoc     GCCCA.CCCGGUcACAGUGAGCGGgcaacAC.CCGGacuCAUUUcgaacCCGGaaGUuaagCCG
Wsp.hung     .uCAAUAGCGGCcACAgcAGGUGUgucacAC.CCGUUCcCAUUCcGAAcaCGGaaGUuaagACA
Paraco       .GUC.UGGUGGccAAAGCACGaGCAaaacAC.CCGAUCcCAUCCcGAacUCGGccGUuaagUGC
B.stearo     ..CC.UAGUGACaAUAGCGGAGAGgaaacAC.CCGUUCcCAUCCcGAacACGGaaGUuaagCUC
B.acidoc     ..UC.UGGUGACuAUAGCGGAGGggcaaCAC.CCGUaccCAUCCcgaacACGGacGUGaagaCC
Strept.      ..UG.UGGUGGCGAUAGCGAGAAGgauacAC.CUGUaacCAUGCcgaacACAGaaGUuaagCUU
B.brevis     ..UC.UGGUGAUGAUGGCGGAGGGgacacAC.CCGUUCcCAUACcGAacACGGccGUuaagCCC
C.pasteu     ..UC.CAGUGUCuAUGACUUAGAGguaacAC.UCCUUCcCAUUCcGAacAGGcaGGUuaagCUC
B.sub        ..UU.UGGUGGCGAUAGCGAAGAGgucacAC.CCGUUCcCAUACcGAacACGGaaGUuaagCUC
B.lichen     ..UU.UGGUGGCGAUAGCGAAGAGgucacAC.CCGUUCcCAUGCcGAacACGGaaGUuaagCUC

                                    C                      C′               A′
E.coli      CGUaGC..gCCgAUgg.uaGUGUGG.GGUCU.CCCCAUGCgaga.GUaGGg.aaCUGCCA.GGCAU
Photobac    AAUaGC..gCCgAUgg.uaGUGUGG.GGUCU.CCCCAUGUgaga.GUaGGa.caUCGCCA.GGCAU
Beneckia    AUUaGC..gCCgAUgg.uaGUGUGG.GGCUU.CCCCAUGUgaga.GUaGGa.caUCGCCA.GGCuU
P. fluor    UGCaUC..GCCgAUgg.uaGUGUGG.GGUUU.CCCCAUGUgaga.GUaGGU.caUCGUCA.AGAuU
Azo.vine    CGCaUC..GCCgAUgg.uaGUGUGG.GGUUU.CCCCAUGUaaga.GUaGGU.cAUCGUCA.GGCGC
P.aerugi    CGCaUC..GCCgAUgg.uaGUGUGG.GGUCU.CCCCAUGUgaga.GUaGGU.cAUCGUCA.AGCuC
R.rubrum    CCCuGC..GCCaAUgg.uaCUGCGU..CUUAAGGCG.UGGgaga.GUaGGU.cGCCGCCA.GGCCU
Wicrococ    CAUaGC..GCCgAUgguuaCUGUAA.CCGGGAGGUUGUGGgaga.GUaGGU.cGCCGCCG.UGA..
Streptom    UACaGC..gCCgAUgg.uaCUGCAG.GGGGGACCCUGUGGgaga.GUaGGa.cGCCGCCG.AAc.U
Th.aquat    GCCaGC..GCCgAUgg.uaCUGGGA.CCGCAGGGUCCUGGgaga.GUaGGU.cGGUGCGGGGGAU
Th.therm    GCCaGC..GCCgAUgg.uaCUGGGC.GGGCGACCGCCUGGgaga.GUaGGU.cGGUGCGGGGGAU
Prochlor    uGCuGC..GGCUA.agauaCUuGCU.GGGUUGCUGGCuGGgaaa.aUAGCU.cGAUGCCA.GGAuU
Hb.salin    GCCuGCGUuCCGGucaguaCUGGAGuGCGCGAGCCUCUGGgaaa.uCCGGuUcGCCGCCU...acU
Hc.morrh    GCCaGCGUuCCAGcgaguaCUGGAGuGUGCGAACCUCUGGgaaa.aCUGGuUcGCCGCCU...ccC
T.acidop    GCU.GCGUAUUGCGUUguaCUGUAUgCCGCGAGGGUACGGgaAGC.GCAAUAuGCUGUUaCC...ACU
S.acidoc    CUC.ACG.UUAGUGGGgccGUGGAUaCCGUGAGGAUCCGCagCCCcACUAA..GCUGGGaUGGGUUUU
Wsp.hung    CCUcacGUGGAUGAcgguaCUGAGGuACGCGAGUCCUCGGgaaa.UCAUCCUcGCUGCUAUUGu.U
Paraco      CGUaGC..GCCaAUgg.uaCU.GCG.UCAAAAGACGU.GGgaga.GUaGGU.caCCGCCA.GAC.C
B.stearo    UCCaGC..GCCgAUgg.uaGUuGGG.GCCAGCGCCCCuGCaaga.GUaGGU.cGUUGCUA.GG..C
B.acidoc    UCCaGC..GCCgagaa.uaCUGGGA.GGGCAGCCUCCUGGgaaa.guaGGU.cGUUGCCA.GG..C
Strept.     CUUaGC..gCCgAUUg.uaGUGAGG.GGGUUGCCCCUUGUgagA.GUaGGa.CGUCGCCA.CG..C
B.brevis    UCCaGC..gCCaAUgg.uaCUuGCU.CCGCAGGGAGCcGGgaga.GUaGGa.CGUCGCCA.GG..C
C.pasteu    UAAuGU..GCUgAUgg.uaCUGCAG.GGGAAGCCCUGUGGaaga.GUaGGU.cGACGCUG.GG..U
B.sub       UUCaGC..gCCgAUgg.uaGUcGGG.GG.UUUCCCCuGUgaga.GUaGGa.CGCCGCCA.AG..C
B.lichen    UUCaGC..gCCgAUgg.uaGUuGGG.GG.CUUCCCCuGUgaga.GUaGGa.CGCCGCCA.AG..C
```

b

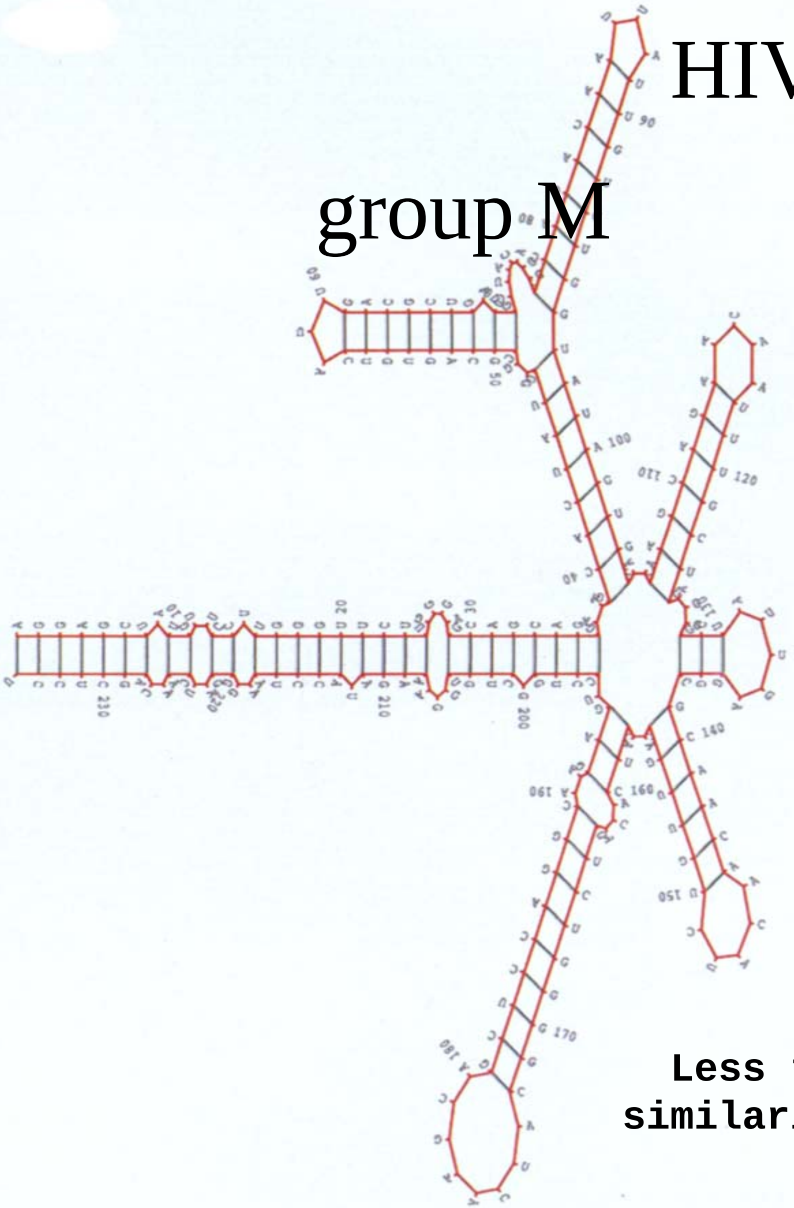
E. Coli 5S rRNA



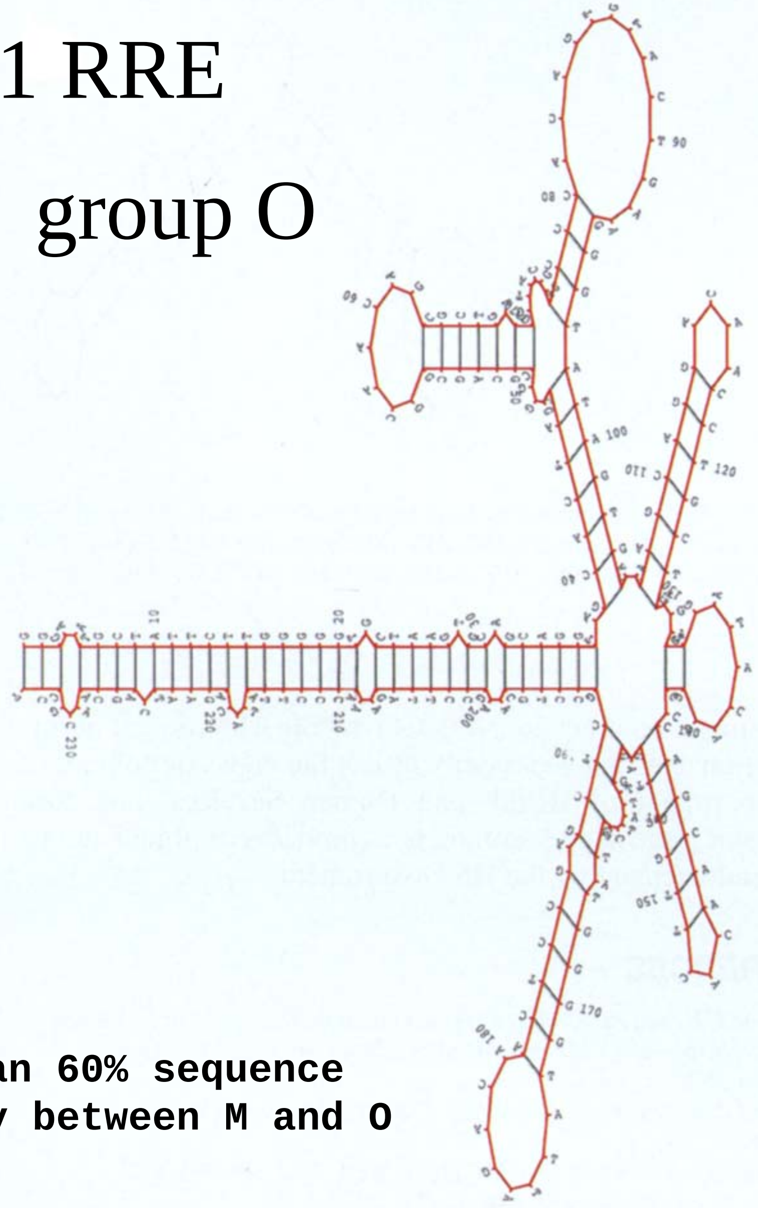
)

HIV-1 RRE

group M



group O



Less than 60% sequence similarity between M and O

Structural alignment of seven HIV-1 RRE sequences

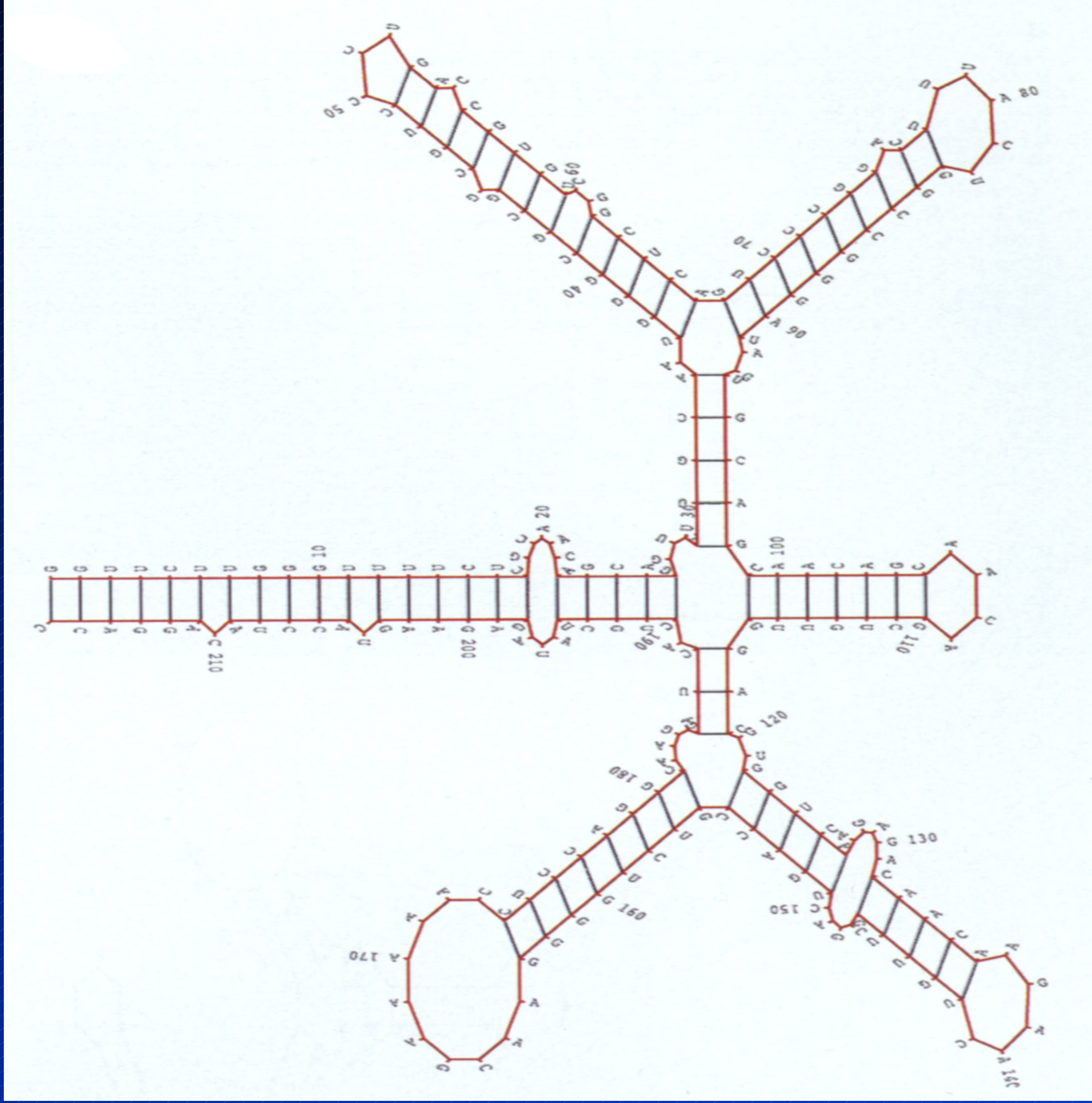
```
                                A                            B1          B2
SF2        AGGA..GCUaUGuuCCuUG.GG.UUCUUgggaGCAGCAGGAAGCACUAUGGGCGCAGUG.UCAUU
HXB2       AGGA..GCUUUGuUCCUUG.GG.UUCUUgggaGCAGCAGGAAGCACUAUGGGCGCAGCc.UCAAU
MAL        AGGA..GCCaUGuUCCUUG.GG.UUCUUgggaGCAGCAGGAAGCACGAUGGGCGCAGCG.UcACU
ELI        AGGA..GCUauGUUCcuUG.GG.UUCUUgggaGCAGCAGGAAGCACgAUGGGCGCA.CGgUCAGU
HIVU455    AGGA..GCUaucUUCcuUG.GG.UUCUUaggaGCAGCuGGAAGCACaAUGGGCGCGGCG.UcAAU
HIVANT70   GGGAaUGCUA...UUCUUGGGGgUUCUAAGu.GCAGCAGGUAGCACUAUGGGCGCAGCG.gcAAC
MVP5180    GGGaaUGCUA...UUCUUGGGGUgCUAAGu.GCaGCAGGUAGCACUAUGGGCGCAGCG.gcAAC

              B2'            B3        B3'      B1'       C
SF2        GACGCUGaCGGUACAGGCCAGACAAUUAUUGUCUGG UAUAGUGCAACAGCAGAACAAUUU
HXB2       GAcGCUGaCGGUACAGGCCAGACAAUUAUUGUCUGG UAUAGUGCAGCAGCAGAACAAUUU
MAL        aACGCUGaCGGUACAGGCCAGACAGUUACUGUCUGG UAUaGUGCAACAGCAAAACAAUUU
ELI        GACGcUGaCGGUACAGGCCAGACAaUUAaUGUCUGG UAUaGUGCAACAGCAAAACAAUUU
HIVU455    aACGCUGaCGGUACAGGCCAGACAAUUAUUGUCUGG UAUaGUGCAACAGCAGAGCAAUCU
HIVANT70   aaCGCUGgCGGUACAGACCCaCAcuUUGcUGaaGGG UAUAGUGCAAcAGCAGGACAACCU
MVP5180    agCGCUGaCGGUACGGACCCaCAGuGUACUGaaGGG UAUAGUGCAACAGCAGGACAACCU

           C'    D     D'   E         E'            F
SF2        GCUGAGGGCUAUUGAGGC GCAACAaCAUcUGUUGCAA..CUCacaGUCUGGGGCAUCAAGCA
HXB2       GCUGAGGGCUAUUGAGGC GCAACAGCAUCUGUUGCAA..cUCacaGUcUGGGGCAUCAAGCA
MAL        GCUGAGGGCUAUAGAGGC GCAACAGCAUCUGUUGCAA..CUCacgGUCUGGGGCAUUAAACA
ELI        GCUGAGGGCUAUAGAGGC GCAACAGCAUCUGUUGCAA..cUCacgGUCUGGGGCAUUAAACA
HIVU455    GCUGAGGGCUAUAGAGGC UCAACAGCAUCUGUUGAAA..CUCacUGUCUGGGGCAUUAAACA
HIVANT70   GCUaAGAGCaAUACAgGC cCAGCAGCAAUUGCUGa..GGCUaucuxuaUGGGGUAUCAGACA
MVP5180    GCUGAGAGCgAUACAgGC cCAGCAACACUUGCUGa..GGuuaucuGUaUGGGGUAUUAGACA

                F'                        A'
SF2        GCUCCAGGCaaGAG UCCUgGCU.GUggaAAGAuA.CC.UAagGGauCAacAGC...UCCU
HXB2       GCUCCAaGCaaGAa UCCUaGCU.GUggaAAGAuA.CC.UAaAGGAuCAAcAGC...UCCU
MAL        GCUCCAGGCaaGAG UCCUgGCU.GUggaAAGAuA.CC.UAaAGGAuCAacGGC...UCCU
ELI        GCUCCAGGCaaGAa UCCUgGCU.GUggaAAGAuA.CC.UAaaGGAUcaacAGC...UCCU
HIVU455    GCUCCAGGCAaGAG UCCugGCU.GUggaAAGAuA.CC.UAcaGGAucaacAGC...UCCU
HIVANT70   ACUCCGAGCUc... GCCU.GCUaGC.CUUAGA.AaCCUUAcuacAGAAUcAGCAacUCCU
MVP5180    ACUCCGaGCUC... GCCU.GCaaGC.CUUAGaaACCCUUA.uacAGAAUcAGCAacgCCU
```

Consensus RRE for HIV-2/SIV



Structural alignment of 10 RRE sequences of HIV-2/SIV

```
                              A                          B1           B2                    B2′
HIV2ROD      GGUUCUUGGGUUUUCUCgcaacAGCAGGUUCUGCAAUGGGCGCGgCGUCCCUGAcCGUGU
SIVMM251     GGUUCUUGGGUUUUCUCGcaacGGCAGGUUCUGCAAUGGGCGCGGCCUcGUUcAGGCUGa
SIVMM142     GGUUCUUGGGUUUUCUCGcaacGGCAGGUUCUGCAAUGGGCGCGgCGUCGUUGAcCGUGa
HIV2BEN      GGUuCUUGGGUUUUCUCGcgacAGCAGGUUCUGCAAUGGGCGCGCGgUCCCUGACGCuGU
HIV2CAM2     GGUUCUUGGGUUUUCUCacaacAGCAGGAGUUGCAAUGGGCACGGCGUCCUUGACGCUGU
HIV2D194     GGUUCUUGGGUUUUCUCGcgacAGCAGGUUCUGCAAUGGGCGGCGCGUCCUUGACGCuGU
HIV2ISY      GGUUcCUAGGUUUUCUCAcgacAGCAGGUGCUGCAAUGGGGGCGGCGUCUCUGACGCUGU
HIV2NIHZ     GGUUCcuAGGUUUUCUCgcaacAGCAGGUUCUGCcAUGGGCGCGGCGUCCUUGACGCUGU
HIV2UC1      GGUUCUUGGGacUUCUUGcaaUGGCAGGUUCUGCAAUGGGCGCAaCGUCCUUGACGcUGU
HIV2ST       GGUUCUUAGGUUUUCUCacgacAGCAGGAGCUGCAAUGGGCGCGGCGUCCUUGACGCUGU

                        B3                B3′          B1′      C              C′    D1
HIV2ROD      cgGCUCA GUCCCGGaCUUUACuGGCCGGGAUAGUGCAG CAACAGCAACAGCUGUUG GACG
SIVMM251     cCGCUCA GUCCCGGaCUUUAUuGGCUGGGAUAGUGCAG CAACAGCAACAGCUGUUG GGCG
SIVMM142     cCGCUCA GUCCCGGaCUUUAUuGGCUGGGAUAGUGCAG CAACAGCAACAGCUGUUG GACG
HIV2BEN      caGCCCA GUCCCGGaCUUUACuGGCCGGGAUAGUGCAG CAACAGCAACAGCUGUUG GACG
HIV2CAM2     caGCCCA GUCUCGGaCUUUAUuGGCCGGGAUAGUGCAG cAACAGCAACAGCUGUUa GACG
HIV2D194     CgGCUCA GUCCCGGaCUUUACuGGCCGGGAUAGUGCAG CAACAGCAACAGCUGUUG GACG
HIV2ISY      CggCUCA GUCUCGGAcuUUAUUCCGuGGgAUAGUGCAG CAACAGCAACAGCUGUUG GACG
HIV2NIHZ     caGCUCA GUCUCGGaCUUUAUuGGCCGGGAUAGuGCAG CAACAGCAACAGCUGUUG GAUG
HIV2UC1      caGCUCA GUCCCGGaCUUUACuGGCUGGGAUAGUGCAG CAGCAGCAACAGCUGCUG GACG
HIV2ST       cgGCUCA GUCUCGGaCUUUAUuGGCCGGGAUAGUGCAG CAACAGCAACAGCUGUUG GACG

                    D2                    D2′         D3                  D3′
HIV2ROD      UGGUCAagagaCAACAaGAAcUGUUGcgacUGACC GUCUGGGGaaCGAAAAacCUCCAGG
SIVMM251     UGGUCAAgagaCAACAAGAAUUGUUGcgaUUGACC GUCUGGGGaaCAAAGAacCUCCAGA
SIVMM142     UGGUCAagagaCAACAAGAAUUGUUGcgacUGACC GUCUGGGGaaCAAAGAacCUCCAGA
HIV2BEN      UaGUCAagagaCAACAaGAAaUGUUGcgacUGACc GUCUGGGGaaCGAAAAacCUCCAGG
HIV2CAM2     UGGUCAagagaCAACAAGAAUUGUUGcgacUGACC GUCUGGGGAACAAAAAUUCUCCAGG
HIV2D194     UGGUCAAgagaCAACAaGAAaUGUUGcgaUUGACC GUCUGGGGAaCGAAAAaUCUCCAGG
HIV2ISY      UGGUCAagagaCAACAaGAAaUGUUGcgacUGACC GUCUGGGGaaCUAAAAacCUCCAGG
HIV2NIHZ     UGGUCAagagaCAACAaGAAaUGUUGcgacUGACC GUCUGGGGAaCAAAAAaUCUCCAGG
HIV2UC1      UGGUCAaaagaCAACAGGAACUGUUGcggcUGACC GUCUGGGGaaCGAAAAacCUCCAGA
HIV2ST       UGGUCAagagaCAACAaGAAaUGUUGcgacUGACC GUCUGGGGAaCAAAAAaUCUCCAGG

                  D1′              A′
HIV2ROD      CAAGAGUCACUGCUauaGAGAAGuACCUAcAGGACC
SIVMM251     CUAGGGUCACUGCCauCGAGAAGuACUUAgAGGACC
SIVMM142     CUAGGGUCUCUGCCauCGAGAAGuACUUAaAGGACC
HIV2BEN      CAAGAGUCACUGCUauCGAGAAGuACCUAaAGcAUC
HIV2CAM2     CAAGAGUCACUGCUauaGAGAAGuACCUAaAGGAUC
HIV2D194     CAAGAGUCACUGCUauCGAGAAAuACUUAAaGGACC
HIV2ISY      CAAGAGUCACUGCUauUGAGAAGuACCUAGcAGaCC
HIV2NIHZ     CAAGAGUCACUGCUauaGAGAAGuCACUaaaGGACC
HIV2UC1      CAAGAGUCACUGCCAuCGAGAAauaCCUAAaGGACC
HIV2ST       CAAGAGUCACUGCUaucGAGAAAuACUUaAAGGACC
```

Atomic-level structural models

Atomic-level models can be derived by manual model building, or by programs.

We use Hugo Martinez’s program RNA2D3D which literally folds a planar secondary structure model into 3D.

Refinements are done with molecular mechanical/dynamical programsl, mainly using the Kollman lab’s AMBER.